A method to produce metadata describing and assessing the quality of spatial landmark datasets in mountain area

Marie-Dominique van Damme, Ana-Maria Olteanu-Raimond

AGILE: GIScience Series, 3, 17, 2022. https://doi.org/10.5194/agile-giss-3-17-2022
Proceedings of the 25th AGILE Conference on Geographic Information Science, 2022.
Editors: E. Parseliunas, A. Mansourian, P. Partsinevelos, and J. Suziedelyte-Visockiene.
This contribution underwent peer review based on a full paper submission.

A method to produce metadata describing and assessing the quality of spatial landmark datasets in mountain area

Marie-Dominique Van Damme [1,*] and Ana-Maria Olteanu-Raimond [1,*]

[1]LASTIG, Univ Gustave Eiffel, IGN-ENSG, 94165-F Saint-Mandé, France
[*]These authors contributed equally to this work.

Correspondence: marie-dominique.van-damme@ensg.eu;ana-maria.raimond@ign.fr

Abstract.

The increase of recreational activities in the mountains and a growing amount of websites proposing geographic data, offer new opportunities for societal needs such as mountain rescue, biodiversity monitoring, outdoor activities. However, the main issue with the websites data is the lack of metadata that minimizes its reuse outside the community that produced the data. The goal of this paper is to study and generate quality and descriptive metadata using ISO standards. To this end, we propose a method based on a common vocabulary such as an ontology and a data matching process. The first one allows to associate to each type of feature from an available geographic dataset an ontology class that will facilitate data matching, reproducibility of results and minimize semantic heterogeneity. The second one allows to define matching links between features representing the same entity in the real world and compute quality indicators based on the validated links. Finally, at the end of this process, we are able to generate descriptive and quality metadata. By following ISO standards and using the QualityML dictionary for measures, the metadata is serialized to XML and can finally be published as open source. Our approach was applied to five different landmark datasets in the French Alps region. New insights were acquired regarding positional accuracy and semantic granularity.

Keywords. Geographic Metadata, Quality Assessment, Geographic OpenData, ISO 19115 and ISO 19157

1 Introduction

Much studies have observed an increase and diversification of outdoor activities in recent years (Routier and Michot, 2021) and the COVID-19 pandemic has significantly changes human mobility patterns with profound changes in recreational use of natural areas (Venter et al., 2020). In this context, geographic data in natural and mountainous areas, which in the past were not considered a priority by national mapping agencies (NMAs), except for making maps, are becoming relevant for different societal needs such as organizing outdoor activities for local stakeholders or citizens (Routier and Michot, 2021) or integrating landmarks from multiple heterogeneous data sources to help rescuers locate victims in mountainous areas (Van Damme et al., 2019).

The growing number of websites offering outdoor leisures provides real opportunities to complement authoritative geographic data to meet societal needs. These websites publish open source data through Application Programming Interfaces (APIs) that are directly accessible from web services. Unless producers create a snapshot of the dataset and publish it on public Geoportals or directly registered as a spatial service on a geographic catalog, which will be rare, these data sources are not well referenced. Furthermore, their documentation does not correspond to the standard description of geographic data sets. In the absence of detailed metadata, searching for and obtaining information about online data remains a difficult task for the user.

The lack of metadata makes it difficult to search, as the data does not respect the findable aspect. Thus, of the fourteen best practices recommended by Brink et al. (2017) for publishing data on the web, the 13th recommends including spatial metadata, where the elements of scope (i.e., spatio-temporal context) and quality are included in the same metadata description report. In addition, an interesting paper describing the history of geographic information standards (Brodeur et al. (2019)) highlights the fact that the scope of the ISO 19115-1:2014 standard contains all the metadata for the description of geographic datasets and that the quality elements have been moved into the ISO 19157:2013 standard. From the standards describing a dataset, while the metadata description summarizes the spatial-temporal context of the spatial datasets, the quality metadata indicates how well the geographic data cor-

respond to the real world, according to the data specifications.

In addition to the findable dimension, metadata is needed for other purposes such as improving the efficiency of the search engine, helping users to evaluate the usability and applicability of a geographic dataset to their needs, etc. For example, based on the existing metadata, some research works have proposed different approaches using keyword enrichment with other vocabularies (Vockner and Mittlböck, 2014), knowledge graph construction (Zrhal et al., 2021) to improve dataset retrieval, keyword semantic similarity calculation (Chen and Yang, 2020) or defining the best web service chaining according to the quality of services (Halilali et al., 2018). Spatial and temporal similarities can be combined with the quality of the metadata input to assign a score to each dataset (Kuo and Chou, 2019). It is worth noting that all these approaches use description metadata.

Specifically, with respect to the usability of geographic datasets, it is expected that quality metadata will be used. The European Inspire Directive provides recommendations on quality indicators to be used in geographical names for example. Many research works follow these recommendations to assess the quality of volunteered geographic information (VGI) by comparing it to reference data ((Zielstra and Zipf, 2010), (Girres and Touya, 2010), (Acheson et al., 2017)). Traditional geographic data quality assessment involves applying a manual (Girres and Touya, 2010) or automatic (Zielstra and Zipf, 2010) data matcging between VGI and reference geographic data, and then calculating quality indicators using on matched featuresISO 19157:2013. Metadata in general, but more specifically quality metadata, can be expressed as an overall metric to quickly tell users whether a dataset meets usability. For example, in France, the National Council for Geographic Information (NCGI) offers a tool to evaluate datasets by assigning stars from 1 to 5 according to the geographic data quality indicators of the ISO 19157:2013. Metadata visualisation solutions such as those proposed by the GeoViQua project [1] for sensor data (Nüst and Lush (2019)) or by (Figgemeier et al. (2021)) to visualise the provenance metadata of a dataset, a complementary information to quality, in the form of a graph, also help users to measure the usability of a dataset.

To apply both the search engine approach and the usability/applicability approach, quality metadata must be field in and expressed in a quantitative form, which is rarely the case. Thus, a first challenge concerning metadata is its availability. Indeed, in the research literature dealing with geographic data quality assessment, indicators are expressed in a quantitative form but rarely described by metadata standards. The study conducted by (Ureña-Cámara et al., 2019) analyzed the metadata records of the Spanish data portal. They found that all metadata records are about metadata description, with no quality metadata reports available to users. Thus, for 3640 geographic metadata dataset records, 97% have a good bounding box and 41% have topological contradictions due to human typing errors. A second challenge is time-consuming metadata editing. For example, (Ureña-Cámara et al., 2019) mentioned that, most of the metadata in a studied space portal is edited manually. Alternatives are available using crowdsourcing communities. For example, (Kalantari et al., 2014) propose a prototype where volunteers can create metadata by sharing their notes and analyzing the keywords used in the user's search. Finally, we found that most of the quality metadata is for image datasets (Figgemeier et al. (2021), Wagner et al. (2021)).

In this context characterized by the lack of existing metadata for vector datasets, our paper contributes to this gap by proposing a method to define and provide description and geographic quality metadata. We study five heterogeneous data sources, mostly produced collaboratively or from authoritative institutions. Furthermore, by analyzing geographic quality indicators, our paper provides new insights regarding the semantic and thematic granularity of the studied data sources in open areas as well as their complementarity and redundancy. In addition, thanks to the definition of the scope of data quality, on-demand metadata focused on specific themes is possible. This allows to focus on certain themes of the data sources such as landform or isolated accommodation. The produced metadata are used to semantically integrate landmark data, coming from heterogeneous and multi-source open data sources, to define a landmark data warehouse for mountain rescue (Van Damme et al., 2019), and more generally can address other societal needs.

The reminder of this paper is organized as follows. Our global approach is defined in Section 2. Section 3 describes the datasets and the main processes to access them. Section 4 describes the implementation of quality metadata for the datasets. Finally, a summary of our contributions and future work is presented in Section 5.

2 Global approach and material for publishing metadata

In order to study the relevance of different data sources in a homogeneous way and to define their characteristics, the idea is to analyze a wide range of aspects of the metadata provided in a serialized format. The approach we proposed for the construction of the metadata is composed of three general steps presented in Figure 1.

Step I identifies outdoor activity data sources using their APIs and downloads the data. Step II is to define and build metadata for the data sources. Two types of metadata are identified: (i) description metadata and (ii) quality metadata. Finally, the third step is to automatically generate XML metadata. This section describes the three steps.

[1] http://www.geolabel.info/

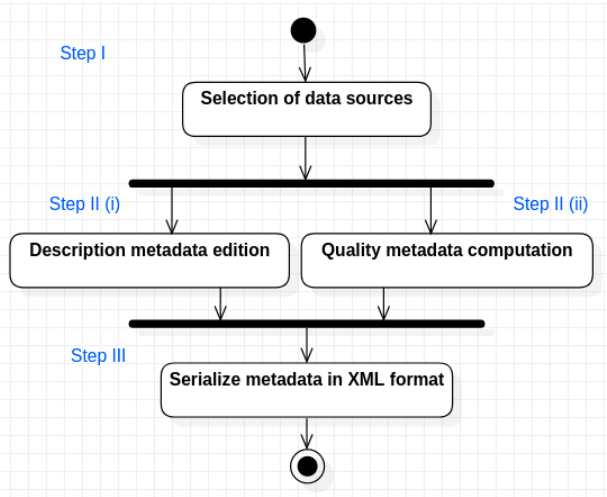

Figure 1. Approach to create metadata records from geographic data sources

2.1 Selection of data sources

The first step in the process is to identify and access potential data sources to collect. The target data warehouse for mountain rescue should look like a gazetteer. Thus, the landmarks should have a point geometry, a type related to the spatial data theme classification in the sources, and a name if one exists. Regarding the data sources, we focus on active websites dedicated to mountain leisure. Five geographic data sources are considered, a combination of authoritative and crowdsourced sources, specialized on a thematic area and of general interest use. The selected data sources, with the exception of OpenStreetMap, are described in detail in (Van Damme et al., 2019), so their description will be very short.

Authoritative data sources. The French NMA provides Points of Interest or Activity themes (relief, building, hydrography, etc.) with a national coverage. These data are part of the national topographic data, named BDTOPO. The data is derived from a shapefile, although there is opendata since January 2021. Protected areas (PAs) run by French public institutions with the role of managing the recreational use of nature is the second authoritative data source, with tourism as the main thematic use. The data is provided through an API in JSON format.

Crowdsourced data sources. Camptocamp (C2C) is a website dedicated to more or less experienced mountaineers. The data concerns topographic landmarks for leisure activities (running, cycling, climbing, etc.). Refuges.info, as its name suggests, provides detailed information about shelters (e.g. name, opening hours, number of places). Other types of features are also provided such as water points, peaks, etc. OpenStreetMap (OSM) is a well-known collaborative project offering many types of topographic and thematic geographic data. The proposed method to build the OSM dataset is detailed in section 3.1. All data were downloaded from an API endpoint.

2.2 Metadata definition

To define the description and quality metadata of the datasets selected in our study, we used two ISO standards, ISO 19115-1:2014 and ISO 19157:2013. These standards are recommended by the INSPIRE Directive for the dissemination of spatial data and the reporting of data quality.

2.2.1 Description metadata edition

As recommended in Brink et al. (2017), the goal of Step II is to define all descriptive metadata using the ISO 19115-1:2014 standard. The role of this metadata is to describe general characteristics such as spatio-temporal context, topic category, resolution, spatial representation, geometry type, etc. We have briefly defined some of the components, but for more details, we invite the reader to refer to the ISO 19115-1:2014 standard.

- *MD_ReferenceSystem*: spatial reference system used.
- *MD_LegalConstraints*: licenses in the dataset
- *EX_GeographicExtent*: the spatial area of the dataset

We can notice that in our study, the high level of metadata in the ISO 19115-1:2014 standard of the chosen spatial datasets, although very useful to exchange spatial data, are not discriminating. Indeed, the spatial representation, reference system, constraint information or identification are very similar from one source to another.

Although the provenance information of a dataset expressed by the lineage element is important, we did not address this aspect for the sake of homogeneity between sources. Indeed, in OSM, some tags indicate the source from which the data comes, but for other sources, neither the attributes nor the documentation allow to know it.

2.2.2 Quality metadata computation

The ISO 19157:2013 standard defines spatial data quality and specifically the components and indicators for describing data quality as well as the metrics for assessing data quality. In contrast to the description metadata, the computations of the quality measures are almost fully automatic in our work. As shown in Figure 2, the measures are computed by comparing the dataset to be evaluated (denoted *DS_Eval*) with a reference dataset of well-known quality (denoted DS_Ref). Our approach for computing data quality measures is inspired by the approach described in (Van Damme et al., 2019).

The first step is the *Semantic Mapping*. This involves aligning dataset element types and classes defined in an application ontology defined for mountain rescue purposes, named Landmarks Ontology. (OOR)[2]. The alignment consists in manually assigning to each feature type of both *DS_Eval* and *DS_Ref* the URI of the corresponding class defined in OOR. The second step automatically

[2] choucas.ign.fr/doc/ontologies/oor.owl/

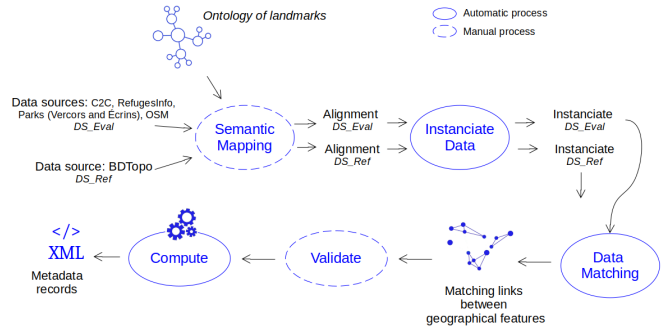

Figure 2. Overview to create quality metadata

instantiates schemas based on alignments and datasets. These steps are new comparing with the approach proposed by (Van Damme et al., 2019) and ensure the reproductibility of our research.

The third step consists in matching each available dataset, DS_Eval_i, to the same reference dataset, *DS_Ref*. The data matching algorithm proposed in Olteanu Raimond et al. (2015) with the same parameters described in (Van Damme et al., 2019) is used. The result of data matching is a set of links with a cardinalityof $1:1$ representing homologous features in both *DS_Eval* and *DS_Ref* datasets.

Finally, the two last steps of the approach are, *Validate* and *Compute*. The first step is to manually validate the matching links to avoid computing quality measures on false homologous features. The last step automatically computes the quality measures listed in the first three columns of the Table 1 based on the true matching links and generates metadata reports. These last two steps are also new compared to the work described in (Van Damme et al., 2019).

2.3 Serialize metadata in XML format

The goal of Step I and Step II is to define metadata by using the two standards ISO 19115-1:2014 (resp. ISO 19157:2013) for the description (resp. for the quality) of datasets. In Step III of the process, quality scope (geographic and temporal extents and the subset of data, if necessary), data quality elements, data quality measures and the processes used to compute data quality (worflow defined in 2.2.2) are encoded. To this purpose, a third standard ISO 19157-2:2016 which specifies how to encode each quality assessment element is used.

If the metadata records are published on a catalog portal, then the measures are usually defined in the registry. The French geocatalog proposes the QuaDoGeo measures. In our case, metadata records will be published in a open-access repository, so we chose the QualityML[3] profile as dictionary to encode our quality metadata as it best meets our needs. One more reason for choosing the standard *QualityML* is that it offers metrics adapted to confusion matrix such as the *Overall accuracy*.

Among the proposed measures listed in (fourth column in Table 1), ten of them are deeply defined in the QualityML dictionary. We explain below two of them that are suitable for our purpose.

Each quality data item may have an indicator that measures the confidence obtained from the matching results *DQ_Confidence*. It is defined as the rate of features for which the algorithm could not make a decision (e.g. undecided cases) and those that were wrongly matched (e.g. the matching link is false without knowing which is the homologous feature in the reference dataset *DS_Ref*) relative to the total number of features to be matched. This indicator is only specified if the computation requires the $1:1$ data matching links.

Since ISO 19157:2013 does not have appropriate measures for assessing the accuracy of feature names, we propose a specific measure used in the toponymic criteria of the data matching. The definition of the Samal distance (Samal et al., 2004) measuring the similarity between the names of matched features is added in the metadata, as a user-defined data quality measure.

Finally, we choose the Java library *Apache SIS library*[4] to generate the xml files. Not all serializations of measures are implemented yet.

2.4 Software and Data Availability

2.4.1 Methods and code availability

The data matching algorithm is coded in java and is available on github: https://github.com/umrlastig/MultiCriteriaMatching

Data quality measures computation. The script to compute data quality measures is coded in SQL and can be accessed from github: https://github.com/ANRChoucas/QualityMetadataSpatialLandmarkDataset.

The *VisuValideMultiCriteriaMatching* tool that allowed us to validate links of data matching is published as an open-source project on github (https://github.com/ANRChoucas/VisuValideMultiCriteriaMatching) under the MIT License.

Metadata file generation. The code allowing to automatically generate a part of the metadata is coded in java and can be find on github: https://github.com/ANRChoucas/QualityMetadataSpatialLandmarkDataset

2.4.2 Data sets availability

Different datasets are published on Zenodo platform[5]. For the sake of clarity, they are listed here.

[3] https://www.qualityml.org/

[4] https://sis.apache.org/

[5] https://zenodo.org/communities/choucasproject/

Table 1. Selected components for metadata quality (see ISO 19157:2013 for more details)

Data quality element			Measure
Class	**Indicator**	**Definition**	**Metric**
Metaquality	Confidence	Trustworthiness of a data quality result	Confidence
PositionalAccuracy	AbsoluteExternalPositionalAccuracy	Closeness of coordinate values to values accepted as or being true	MeanAbsolute2D, RootMeanSquareError, Agreement Rate
ThematicAccuracy	NonQuantitativeAttributeAccuracy	Accuracy of non-quantitative attributes	Samal string similarity
ThematicAccuracy	Thematic ClassificationCorrectness	Comparison of the classes assigned to features to the reference dataset	OverAllAccuracy, ConfusionMatrix
Completeness	CompletenessOmission	Data absent from the dataset, as described by the specifications	MissingItems
Completeness	CompletenessOmission	Classes absent from the dataset, as described by the scope	MissingClass
Completeness	CompletenessComission	Excess data present in the dataset, as described in the specifications	Excess
Completeness	CompletenessComission	Duplicate data	Duplicate

- *landmark datasets*: the five spatial landmark datasets can be access here: https://doi.org/10.5281/zenodo.6480985
- *alignment files*: the five csv files representing alignments between the type of landmarks and a commom vocabulary extracted from an mountain rescue application ontology, named Ontology of landmarks (OOR) can be found here: https://doi.org/10.5281/zenodo.6481338
- *matching results*: the four files representing validated and no validated matching links, the non-matched landmarks and uncertain links are available at: https://doi.org/10.5281/zenodo.6483784
- *metadata*: represents descriptive and quality metadata. Four files represent dataset description with the ISO Standard and one file represent the description of the Samal measure. There are available at: https://doi.org/10.5281/zenodo.6494267

To facilitate the reproducibility of our work, we define two user stories.

User story 1: John is a PhD student working on data quality and needs to learn about data quality assessment. For that, he wants to replicate our methods and tools to recalculate data quality indicators and, at the end, generate the metadata files. To do this, he needs to follow the steps below: (1) download the result link files from Zenodo platform; (2) import the matching results into a database with the model described in the readme and run sql script to compute the quality indicators; (3) run the java *MainMetadataChoucas.java* file that will generate the metadata files.

User story 2: Alice is a researcher and plans to follow our approach to produce metadata for the same landmark datasets but in different areas. To this end, she needs to follow the steps below: (1) download the datasets from the API. The url are specified in the metadata description files; (2) Once data downloaded, she needs to instanciate the database with the alignment files;(3) run the data matching algorithm to match each dataset to be analysed with the reference dataset; (4) validate the matching links by using the VisuValideMultiCriteriaMatching tool and import them into a database (we used PostgreSQL database); (5) run sql script to compute the quality indicators; (6) run *MainMetadataChoucas.java* that will generate the quality metadata; (7) upgrade in a new version the description metadata files on Zenodo and modify them according to the new study area by adding a new report.

3 Selecting, mapping and describing datasets

As one of the objectives is to bring new insights and generate semantic knowledge about open spatial data for mountain rescue, and more generally for park management or outdoor activities, our study area is located in mountains.

3.1 The corpus of dataset

In this study, we focused on areas that do not contain many urban districts but mainly sport practice areas to ensure the same spatial and thematic coverage for all data sources (see Figure, 3). Features are mostly coming from OSM (41,454) and BDTOPO (17,769) in contrast to the thematic datasets: C2C (2,289), Refuges.info (659) and PA (1,906).

Concerning C2C, Refuges.info, PA and BDTOPO no transformation on the data has been performed. Only many requests to the API or reading the files, without filtering, allowed to create datasets. All collected spatial data were transformed to Lambert93 projection (SRID: 2154). Start-

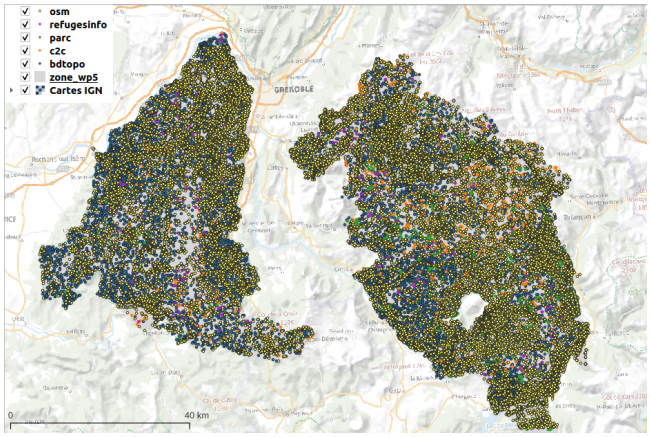

Figure 3. Five data sources for test area located in the French Alps

ing from the work of Gendner et al. (2021), the alignment of all feature types was done manually. For BDTOPO source, another attribute, *detailled nature*, complete the type and nature of features to align them to the best OOR class.

On the other hand, for OSM data, a subset of all collected data is generate. First, data is downloaded from OverpassApi[6] by specifying the test area bounding box. Then, the alignment between the OSM tags and the OOR classes is defined manually by using OSM wiki which defines each tag presented in OSM. Note that only OOR classes corresponding to the mountain outdoor theme such as building theme, natural elements, leisure infrastructure, hydrographic, transportation, and tourist information were instantiated using the OSM data. After transforming spatial data into Lambert93 projection (SRID: 2154) the centroid calculation was applied on the polygonal and linear geometries in order to build the point geometries.

At the end of this step, collected data are distributed as follow: 102 classes are instantiated for the BDTOPO dataset, 123 classes for OSM, 7 for PA, 6 for Refuges.info, and 17 for C2C. This approach provides an unified vocabulary for the data matching process.

Overall, one feature type in a dataset corresponds to one class in the OOR. The exception is mostly for housing. For example, in C2C the type "shelter" is dissociated from "bivouac" ans "shelter". In Refuges.info, "mountain building" and "lodge" are associated with "accommodation building" because the definition in the specification does not match to the definition of the more detailed OOR class.

3.2 Results for description metadata

This section presents metadata that can be defined through website documentation or automatically computed and do not require spatial data matching tools. Among the concepts defined by ISO 19115-1:2014, we focus on four of them to describe the data sources studied in our work.

Licences of dataset are defined in *MD_LegalConstraints* and as shown in Table 2, they are different from one source to another. The CC-BY-SA licenses are more restrictive because the redistribute of the work made from the database must be free (contaminating aspect of the license), whereas with the open licenses, the only condition to redistribute the data is to mention who is the producer of the dataset.

Table 2. License of the five dataset

Class	Indicator
Refuges.info	CC-By-Sa 2.0
C2C	CC-by-nc-nd
PA	Etalab 2.0
OSM	Open Data Commons Open Database License (ODbL)
BDTOPO	Etalab 2.0

The content of *MD_ReferenceSystem* metadata corresponds to the projection Lambert93 (EPSG:2154) following the transformations performed on the data. The *EX_Extent.EX_GeographicExtent* of each dataset is replaced by the bounding box of the test area even though some data sources cover a larger area.

4 Quality metadata creation

This section presents the results of the quality metadata and the serialization in XML format.

Two categories of data quality reports were produced. The first contains the metadata defined in Table 1 for each dataset located in the study area. The second category of data quality report focuses on a subset of the dataset defined by the data whose type is grouped in a same branch of *OOR*. This illustrates what we call on-demand metadata. The same quality metadata defined in Table 1 is then computed. For example, two quality reports are generated for "isolated accommodation" containing shelter, hut, etc. and "convex relief" containing concepts such as summit, peak, rock. The *scope* column of the Table 4 indicates the scope: *all* for the first category and *name of OOR class* for the second.

4.1 Data matching results

In the previous section, the data instantiated using an alignment with the OOR are presented. Then, the four datasets, OSM, C2C, PA and RefugesInfo are matched to the same reference dataset (BTOPO). All matching links results are presented in Table 3. Link cardinality $1:0$ means that a landmark from DS_Eval_i (e.g. C2C) has no homologue landmark in *DS_Ref* (BDTOPO) whereas $1:1$ means that it exists. The link tagged «uncertain» characterises complex cases where the data matching algorithm

[6] http://overpass-api.de/

cannot take any decision. For this paper, only the 1 : 1 links are manually validated by experts.

Table 3. Data matching results and confidence for the scope *all*

Matching links		C2C-BDTOPO	PA-BDTOPO	Refuges.info-BDTOPO	OSM-BDTOPO
1:0		404	960	105	22 756
1:1	total	1435	367	453	5919
	including validated	80%	68%	92%	60%
	DQ confidence	82%	85%	82%	93%
Uncertain		121	164	33	4534

Some quality methods need all matching links to compute the indicators, whereas others, such as DQ_PositionalAccuracy and DQ_ThematicAccuracy, only need 1 :1 links; these links are validated in our process. Thus, for these two quality data elements, we can compute a quality assessment expressing the trustworthiness of the data matching result (i.e. *DQ_Confidence*). The measure is defined as the rate of features for which the data matching algorithm could not make a decision (e.g. undecided) and mismatched links (e.g. the link is false but it is not possible to know whether the homologous feature belongs to the reference dataset, BDTOPO) relative to the total number of features to be matched.

The results of *DQ_Confidence* from the different data sources are shown in the Table 3. We observe that the confidence values related to data matching are high for all sources (more than 82% for C2C, PA, Refuges.info, and 93% for OSM). This means that the values of the quality indicators calculated based on the matching links are reliable. Below, an example of how confidence is serialized in XML format (C2C dataset) is illustrated.

```
<gmd:DQ_MetaqualityConfidence>
  <gmd:nameOfMeasure>
    <gco:CharacterString>Confidence</gco:CharacterString>
  </gmd:nameOfMeasure>
  <gmd:result>
    <gmd:DQ_QuantitativeResult>
      <gmd:valueUnit xlink:href="http://www.opengis.net/def/uom/OGC/1.0/
          ↪ unit" />
      <gmd:value>
        <gco:Record xsi:type="xs:double">0.85</gco:Record>
      </gmd:value>
    </gmd:DQ_QuantitativeResult>
  </gmd:result>
</gmd:DQ_MetaqualityConfidence>
```

The second column of the Table 4 named Metaquality shows when *DQ_Confidence* can be computed.

All indicators, except *duplicate* are computed by using the same method, described in Section 2. The computation of the date, the algorithm citation, the description of the method and the type of evaluation (here *direct external*) are indicated in the metadata element *DQ_EvaluationMethod*.

4.2 Results for quality metadata

The data quality components most suitable for the studied datasets with respect to their geometric and thematic characteristics are selected (see Table 1) and the results are presented on Table 4.

AbsoluteExternalPositionalAccuracy.

We remember that to compute the three positional accuracy measures, we operate on validated links. Next, we use the Euclidean distance to measure the planimetric error as the average of the deviations in x and y, the root mean square planimetric error and the level of agreement (i.e. proportion of the number of positions with planimetric error below a threshold with respect to the total number of measured positions). The threshold is the same for all datasets to ensure comparability and is equal to 30m.

Data coming from Refuges.info have a good planimetric accuracy (23.79m) compared to the other sources. If we look the values more closely, we notice differences by type. For example, for the same dataset (Refuges.info), isolated accommodation have a much better accuracy (13.64m) than the landform landmarks (43.27m).

ThematicClassificationCorrectness.

Semantic accuracy is defined using the confusion matrix that compares the OOR ontology classes assigned to the homologous landmarks. The overall accuracy is the number of correctly classified features in the *DS_Eval* divided by the total number of matched features. In our context, due to the difference of granularity, some different types of landmarks are assigned to the same class such as shelter and accommodation building. Another common case is place names. When a place have only one usage like a hut or a pass, the classes are correctly classified as identical.

We can notice that PA source has a very low overall accuracy (0.24) which shows a high heterogeneity in the classification of PA features. Above all, the confusion matrix shows that these two data sources are not comparable because they have different points of view.The same semantic heterogeneity is observed for the OSM data.

NonQuantitativeAttributeCorrectness.

In order to evaluate the correctness of features names, we compute the Samal distance Samal et al. (2004) for each homologous feature. This distance is between 0 and 1, where 0 means that the two homologous features have exactly the same name and 1 means that the two features have very different names.

Figure, 4 represents the distribution of the values of Samal distance for names in Refuges.info data source. It can be observed that for Refuges.info data source the median is equal to 0.1 and the outliers are below 0.25.

Note that the score are fairly good, but that the level of agreement (threshold is 0.1) shows variation in the names of the features. This diversity of names is an asset and a

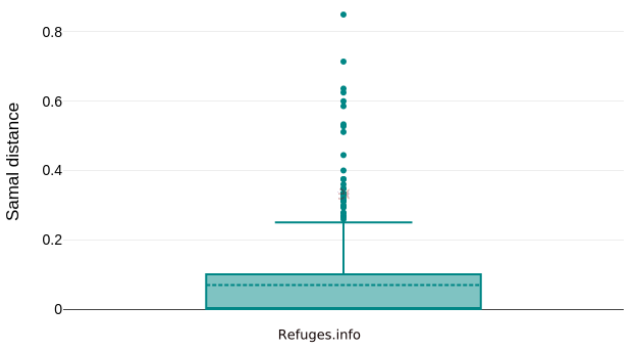

Figure 4. Boxplot of Samal distance for names in Refuges.info source

valuable aid for place names research. The example below shows how normal distribution measure is implemented.

```
<gmd:result>
  <gmd:DQ_QuantitativeResult>
    <gmd:valueType>
      <gco:RecordType xlink:href="http://www.uncertml.org/distributions/
          ↪ normal">Value for string distances</gco:RecordType>
    </gmd:valueType>
    <gmd:valueUnit xlink:href="http://www.opengis.net/def/uom/OGC/1.0/unit"
        ↪ />
    <gmd:value>
      <gco:Record>
        <un:NormalDistribution>
          <un:mean>0.02</un:mean>
          <un:variance>0.13</un:variance>
        </un:NormalDistribution>
      </gco:Record>
    </gmd:value>
  </gmd:DQ_QuantitativeResult>
</gmd:result>
```

CompletenessCommission.

Duplicate is a metric that measures how many features with different identified have exactly the same geometry. Our results shows that OSM and C2C data sources has mainly duplicate features (77 features, 11 features respectively). It is due to the crowdsourced platform.

The measure *Excess* counts the number of features *DS_Eval* present and absent from *DS_Ref*. We can observe relatively low values of commission (0.39 and 0.3 for C2C and Refuges.info respectively) and high values (0.85) for the PA data sources. This shows that the three sources are complementary to the reference data source BDTOPO.

Note that the *Excess* for the OSM data source is not computed. The reason is related to the selection of the data we made for OSM data source. Indeed, knowing the wide spectrum of data types that can be found in OSM, we only selected features that represent landmarks, while for the other sources we considered all data since all of them are landmarks. Thus, due to this selection, the commission indicator doesn't make sense anymore.

CompletenessOmission. It counts the number of features in *DS_Eval* that are missing compared to the total number of features in *DS_Ref*. The omission is computed by taking into account the $0:1$ links in the output of the data matching process.

Omission has high values for all landmarks types and data sources (0.94, 0.98, and 0.99). This shows that the studied data sources are non-exhaustive compared to BDTOPO. This result is not surprising since they are thematic data sources whose objective is not to represent all the entities of the real world but only those linked to outdoor activities.

As for commission indicator, the omission indicator for OSM data is not computed.

Overall, omission and commission indicators show that data sources studied are complementary and few redundancies exist. This is a relevant result for the integration of multi-source data.

Finally, *Missing class "name"* represents the rate of nameless features. The results show that only the OSM data source has features without name (i.e., 50% of OSM features have no name); the other data sources have maximum values (i.e., 100%).

5 Conclusion

This research work aimed to propose a method to generate metadata for spatial open data sources. To this end, we focused on two types of metadata: description metadata and quality metadata which are defined by using three ISO standards: ISO 19115-1:2014 (for metadata description), ISO 19157:2013 (for quality assessment), and ISO 19157-2:2016 (for encoding the metadata). Description metadata are manually field in by using the description of the data sources or by spatial analysis, whereas the quality metadata are generated semi-automatically by a nearly automatic process based on schema alignment using a common vocabulary, data matching, and indicators computation. An innovative aspect of our approach concerns the definition of metadata on-demand. This is possible by defining a specific scope and thus customizing the use of the metadata. The approach is applied on five data sources containing landmarks in mountain areas: authoritative datasets (BDTOPO, and National and Regional Parcs) and crowdsourced datasets (refuges.info, Camp2Camp, and OpenStreetMap).

Concerning the metadata, our work improves the FAIR (Findable, Accessible, Interoperable and Reusable) principles. For example, the Findable and Interoperable principles are improved by the metadata we have generated for most of the relevant data sources in mountain areas that can be used for different purposes: mountain rescue, ecosystem monitoring, tourism, outdoor activities organization. The detailed description and quality metadata of the data sources help users to better assess the usability of the data sources. The principle of accessibility is enhanced through the publication of our results: a minimum of four reports encoded in xml files, four datasets of data matching links, and five alignment models for each dataset. The alignment models that were manually defined in our work, can be reused to apply our approach on other

sites, for different purposes, or different context like rescue in tropical forest. From this perspective, the alignment models can be considered as ground truth data. Interestingly, with semantic analysis through the confusion matrix, users can easily analyze the granularity of semantics between sources, which can help data integration. Finally, our approach is reproducible, the methods (data matching, quality measures computation, xml generation files), tools (data matching validation) as well as the results are openly available.

From a quality point of view, the quality indicators highlighted the richness and complementarity of the studied datasets as well as the semantic and thematic heterogeneity (different names). The first is an issue for data integration, the second is indeed a valuable asset which would be explored for mountain rescue. New insights were also acquired regarding the position accuracy and semantic granularity.

Concerning the future works, one aspect is to improve the manually tasks. Validation task, which is very time consuming, can be improved by using our validation datasets. Indeed, when analysing the validated matching links we realized that rules can be defined. Thus, by using our validation dataset such as ground truth data, we intend to apply a machine learning algorithm such as RIPPER algorithm to derive rules to automatically validate matching links. For the alignment task, methods exist to align automatically vocabularies by semantic web technologies.

Knowing the semantic granularity between sources, a future work is to define, in addition to the alignment between features, relationships between features (e.g. a ski slope is border by two ski lift stations). This is possible by exploiting the data matching results.

Finally, one major future work we will tackle, is the multi-source data integration to build a graph data warehouse of landmarks and routes for mountain rescue purposes. The produced metadata, data matching results as well as the alignments will be used to define a semantic based multi-source data integration.

6 Acknowledgments

This work was supported by the ANR under grant agreement no. ANR-16-CE23-0018 (CHOUCAS research Project: Heterogeneous data integration and spatial reasoning for localizing victims in mountain area).

References

Acheson, E., De Sabbata, S., and S Purves, R.: A quantitative analysis of global gazetteers: Patterns of coverage for common feature types, Computers Environment and Urban Systems, 64, 309–320, https://doi.org/10.1016/j.compenvurbsys.2017.03.007, 2017.

Brink, L., Barnaghi, P., Tandy, J., Atemezing, G., Atkinson, R., Cochrane, B., Fathy, Y., García Castro, R., Haller, A., Harth, A., Janowicz, K., Kolozali, , Leeuwen, B., Lefrançois, M., Lieberman, J., Perego, A., Phuoc, D., Roberts, B., Taylor, K., and Troncy, R.: Best Practices for Publishing, Retrieving, and Using Spatial Data on the Web, Semantic Web, 10, https://doi.org/10.3233/SW-180305, 2017.

Brodeur, J., Coetzee, S., Danko, D., Garcia, S., and Hjelmager, J.: Geographic Information Metadata—An Outlook from the International Standardization Perspective, ISPRS International Journal of Geo-Information, 8, https://doi.org/10.3390/ijgi8060280, 2019.

Chen, Z. and Yang, Y.: Semantic relatedness algorithm for keyword sets of geographic metadata, Cartography and Geographic Information Science, 47, 125–140, https://doi.org/10.1080/15230406.2019.1647797, 2020.

Figgemeier, H., Henzen, C., and Rümmler, A.: A Geo-Dashboard Concept for the Interactively Linked Visualization of Provenance and Data Quality for Geospatial Datasets, 2021.

Gendner, V., Van Damme, M.-D., and Olteanu-Raimond, A.-M.: Modelling and building of a graph database of multi-source landmarks to help emergency mountain rescuers, in: International Cartographic Association, vol. 3 of *Abstr. Int. Cartogr. Assoc., 3, 90, 2021*, Florence, Italy, https://hal.archives-ouvertes.fr/hal-03484804, 2021.

Girres, J.-F. and Touya, G.: Quality Assessment of the French OpenStreetMap Dataset, Transactions in GIS, 14, 435–459, https://doi.org/https://doi.org/10.1111/j.1467-9671.2010.01203.x, 2010.

Halilali, M. S., Gouarderes, E., Devin, F., and Gaio, M.: Plateforme logicielle pour l'intégration et la composition de services géospatiaux, in: Sageo 2018, Montpellier, France, https://hal-univ-pau.archives-ouvertes.fr/hal-02462322, 2018.

ISO 19115-1:2014: Geographic information - Metadata - Part 1: Fundamentals.

ISO 19157-2:2016: Geographic information - Data quality - Part 2: XML Schema Implementation.

ISO 19157:2013: Geographic information - Data quality.

Kalantari, M., Rajabifard, A., Olfat, H., and Williamson, I.: Geospatial Metadata 2.0 – An approach for Volunteered Geographic Information, Computers, Environment and Urban Systems, 48, 35–48, https://doi.org/10.1016/j.compenvurbsys.2014.06.005, 2014.

Kuo, C.-L. and Chou, H.-C.: Metadata assessment for efficient open data retrieval, in: Accepted Short Papers and Posters from the 22nd AGILE Conference on Geo-Information Science (AGILE 2019), Cyprus, Greece, 2019.

Nüst, D. and Lush, V.: A GEO label for the Sensor Web, https://doi.org/10.31223/osf.io/ka38z, 2019.

Olteanu Raimond, A.-M., Mustiere, S., and Ruas, A.: Knowledge formalization for vector data matching using Belief Theory, JOURNAL OF SPATIAL INFORMATION SCIENCE, in press, https://doi.org/10.5311/JOSIS.2015.10.194, 2015.

Routier, G., L.-B. A. D. and Michot, T.: Sports et loisirs de nature en France / Points de repère et chiffres clés issus du baromètre sport 2018 EN: Sports and nature-based leisure activities in France / Benchmarks and key figures from the 2018 sports barometer, Cartography and Geographic Information Science,

pp. 1–24, https://www.sportsdenature.gouv.fr/publications/sports-et-loisirs-de-nature-en-france-2021, 2021.

Samal, A., Seth, S., and Cueto, K.: A feature-based approach to conflation of geospatial sources, International Journal of Geographical Information Systems, 18, 459–489, https://doi.org/10.1080/13658810410001658076, 2004.

Ureña-Cámara, M. A., Nogueras-Iso, J., Lacasta, J., and Ariza-López, F. J.: A method for checking the quality of geographic metadata based on ISO 19157, International Journal of Geographical Information Science, 33, 1–27, https://doi.org/10.1080/13658816.2018.1515437, 2019.

Van Damme, M.-D., Olteanu-Raimond, A.-M., and Ménéroux, Y.: Potential of crowdsourced data for integrating landmarks and routes for rescue in mountain areas, International Journal of Cartography, 5, 195–213, https://doi.org/10.1080/23729333.2019.1615730, 2019.

Venter, Z. S., Barton, D. N., Gundersen, V., Figari, H., and Nowell, M.: Urban nature in a time of crisis: recreational use of green space increases during the COVID-19 outbreak in Oslo, Norway, Environmental Research Letters, 15, 104 075, https://doi.org/10.1088/1748-9326/abb396, 2020.

Vockner, B. and Mittlböck, M.: Geo-Enrichment and Semantic Enhancement of Metadata Sets to Augment Discovery in Geoportals, ISPRS International Journal of Geo-Information, 3, 345–367, https://doi.org/10.3390/ijgi3010345, 2014.

Wagner, M., Henzen, C., and Müller-Pfefferkorn, R.: A Research Data Infrastructure Component for the Automated Metadata and Data Quality Extraction to Foster the Provision of FAIR Data in Earth System Sciences, AGILE: GIScience Series, 2, 41, https://doi.org/10.5194/agile-giss-2-41-2021, 2021.

Zielstra, D. and Zipf, A.: A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany, 2010.

Zrhal, M., Bucher, B., Van Damme, M.-D., and Hamdi, F.: Spatial Dataset Search: Building a dedicated Knowledge Graph, AGILE: GIScience Series, 2, 43, https://doi.org/10.5194/agile-giss-2-43-2021, 2021.

Table 4. Quality measures for the four datasets

Element	Evaluation method	Metaquality	Measure	C2C	Refuges.info	PA	OSM	Scope
PositionalAccuracy	Method described in 2.2.2	Confidence	**MeanAbsolute2D**	47.05m	23.79m	60.7m	49.22m	*all*
			RootMeanSquareError	70.48	39.05	89.26	76.92	
			AgreementRate threshold = 30 meters	0.49	0.76	0.48	0.52	
			MeanAbsolute2D	17.86	13.64	42.18	19.26	Isolated accommodation
			RootMeanSquareError	29.9	21.76	65.26	48.51	
			MeanAbsolute2D	49.06	43.27	90.87	51.05	Landform
			RootMeanSquareError	71.08	56.54	128.1	75.93	
ThematicClassificationCorrectness	Method described in 2.2.2	Confidence	**Overall accuracy**	0.75	0.65	0.24	0.13	*all*
			Confusion Matrix	[31,31]	[25,25]	[36,36]	[126,126]	
			Overall accuracy	0.76	0.26	0.8	0.8	Isolated accommodation
			Confusion Matrix	[2,2]	[4,4]	[3,3]	[3,3]	
NonQuantitativeAttributeAccuracy	Method described in 2.2.2	Confidence	**Mean Samal Distance**	0.05	0.07	0.1	0.18	*all*
			RootMeanSquareError Samal Distance	0.13	0.15	0.21	0.38	
			Agreement Rate Samal Distance	85%	75%	71%	70%	
			Mean Samal Distance	0.03	0.05	0.05	0.05	Landform
			RootMeanSquareError Samal Distance	0.11	0.13	0.09	0.17	
Completeness	Method described in 2.2.2	–	**Excess**	0.39	0.30	0.85	–	*all*
			Missing items	0.94	0.98	0.99	–	
			Missing class *nom*	1	1	1	0.5	
CompletenessCommission	direct internal	–	**Duplicate**	11	0	1	77	*all*